represented on one or two sheets of paper and has obviated the need for direct computer use in most of the systems. The contributions of clinical algorithms to the distribution and delivery of health care, to the training of paramedics, and to quality care audit, have been impressive and substantial. However, the methodology is not suitable for extension to the complex decision tasks to be discussed in the following sections.

3 Databank Analysis for Prognosis and Therapy Selection

3.1 Overview

Automation of medical record keeping and the development of computer-based patient databanks have been major research concerns since the earliest days of medical computing. Most such systems have attempted to avoid direct interaction between the computer and the physician recording the data, with the systems of Weed [115], [116] and Greenes [32] being notable exceptions. Although the earliest systems were designed merely as record-keeping devices, there have been several recent attempts to create programs that could also provide analyses of the information stored in the computer databank. Some early systems [32], [47] had retrieval modules that identified all patient records matching a Boolean combination of descriptors; however, further analyses of these records for decision making purposes was left to the investigator. Weed has not stressed an analytical component in his automated problem-oriented record [116], but others have developed decision aids which use medical record systems fashioned after his [96].

The systems for databank analysis all depend on the development of a complete and accurate medical record system. If such a system is developed, a number of additional capabilities can be provided: (1) correlations among variables can be calculated, (2) prognostic indicators can be measured, and (3) the response to various therapies can be compared. A physician faced with a complex management decision can look to such a system for assistance in identifying patients in the past who had similar clinical problems and can then see how those patients responded to various therapies. A clinical investigator keeping the records of his study patients on such a system can utilize the program's statistical capabilities for data analysis. Hence, although these applications are inherently data-intensive, the kinds of "knowledge" generated by specialized retrieval and statistical routines can provide valuable

assistance for clinical decision makers. For example, they can help physicians avoid the inherent biases that result when the individual practitioner bases his decisions primarily on his own anecdotal experience with one or two patients having a rare disease or complex of symptoms.

There are many excellent programs in this category, one of which is discussed in some detail in the next section. Several others warrant mention, however. The HELP System at the University of Utah [109], [111], [112] utilizes a large data file on patients in the Latter-Day Saints Hospital. Clinical experts formulate specialized "HELP sectors" which are collections of logical rules that define the criteria for a particular medical decision. These sectors are developed by an interactive process whereby the expert proposes important criteria for a given decision and is provided with actual data regarding that criterion based on relevant patients and controls from the computer databank. The criteria in the sector are thus adjusted by the expert until adequate discrimination is made to justify using the sector's logic as a decision tool[4]. The sectors are then utilized for a variety of tasks throughout the hospital.

Another system of interest is that of Feinstein et al. at Yale [17]. They had specific patient management decisions in mind when they developed their interactive system for estimating prognosis and guiding management in patients with lung cancer. Similarly, Rosati et al. have developed a system at Duke University which utilizes a large databank on patients who have undergone coronary arteriography [82]. New patients can be matched against those in the databank to help determine patient prognosis under a variety of management alternatives.

3.2 Example

One of the most successful projects in this category is the ARAMIS system of Fries [20]. The approach was designed originally for use in an outpatient rheumatology clinic, but then broadened to a general clinical database system (TOD) [118], [119] so that it became transferable to clinics in oncology, metabolic disease, cardiology, endocrinology, and certain pediatric subspecialties. All clinic records are kept in a flow-charting format in which a column in a large table indicates a specific clinic visit and the rows indicate the relevant clinical parameters that are being followed over time.

---

[4]This process might be seen as a tool to assist with the formulation of clinical algorithms as discussed in the previous section. Another approach using databank analysis for algorithm development is described in [26].

These charts are maintained by the physicians seeing the patient in clinic, and the new column of data is later transferred to the computer databank by a transcriptionist; in this way time-oriented data on all patients are kept current. The defined database (clinical parameters to be followed) is determined by clinical experts, and in the case of rheumatic diseases has now been standardized on a national scale [36].

The information in the databank can be utilized to create a prose summary of the patient's current status, and there are graphical capabilities which can plot specific parameters for a patient over time [118]. However, it is in the analysis of stored clinical experience that the system has its greatest potential utility [21]. In addition to performing search and statistical functions such as those developed in databank systems for clinical investigation [45], [59], ARAMIS offers a prognostic analysis for a new patient when a management decision is to be made. Using the consultative services of the Stanford Immunology Division, an individual practitioner may select clinical indices for his patient that he would like matched against other patients in the databank. Based on 2 to 5 such descriptors, the computer locates relevant prior patients and prepares a report outlining their prognosis with respect to a variety of endpoints (e.g., death, development of renal failure, arthritic status, pleurisy, etc.). Therapy recommendations are also generated on the basis of a response index that is calculated for the matched patients. A prose case analysis for the physician's patient can also be generated; this readable document summarizes the relevant data from the databank and explains the basis for the therapeutic recommendation.

The rheumatologic databank generated under ARAMIS has now been expanded to involve a national network of immunologists who are accumulating time-oriented data on their patients. This national project seeks in part to accumulate a large enough databank so that groups of retrieved patients will be sizable and thus control for some observer variability and make the system's recommendations more statistically defensible.

3.3 Discussion of the Methodology

The databank analysis systems described have powerful capabilities to offer to the individual clinical decision maker. Furthermore, medical computing researchers recognize the potential value of large databanks in supporting many of the other decision making approaches discussed in subsequent sections. There

are important additional issues regarding databank systems, however, which are discussed below.

(1) Data acquisition remains a major problem. Many systems have avoided direct physician-computer interaction but have then been faced with the expense and errors of transcription. The developers of one well accepted record system still express their desire to implement a direct interface with the physician for these reasons, although they recognize the difficulties encountered in encouraging hands-on use of a computer system by doctors [100].

(2) Analysis of data in the system can be complicated by missing values that frequently occur, outlying values, and poor reproducibility of data across time and among physicians.

(3) The decision aids provided tend to emphasize patient management rather than diagnosis. Feinstein's system [17] is only useful for patients with lung cancer, for example, and the ARAMIS (TOD) prognostic routines, which are designed for patient management, assume that the patient's rheumatologic diagnosis is already known.

(4) There is no formal correlation between the way expert physicians approach patient management decisions and the way the programs arrive at recommendations. Feinstein and Koss felt that the acceptability of their system would be limited by a purely statistical approach, and they therefore chose to mimic human reasoning processes to a large extent [53], but their approach appears to be an exception.

(5) Data storage space requirements can be large since the decision aids of course require a comprehensive medical record system as a basic component.

Slamecka has distinguished between structured and empirical approaches to clinical consulting systems [96], pointing out that databanks provide a largely empirical basis for advice whereas structured approaches rely on judgmental knowledge elicited from the literature or the minds of experts. It is important to note, however, that judgmental knowledge is itself based on empirical information. Even the expert "intuitions" that many researchers have tried to capture are based on that expert practitioner's own observations and "data collection" over years of experience. Thus one might argue that large, complete, and flexible databanks could form the basis for large amounts of judgmental knowledge that we now have to elicit from other sources. Some researchers have indicated a desire to experiment with methods for the automatic generation of medical decision rules from databanks, and one component of the

research on Slamecka's MARIS system is apparently pointed in that direction [96]. Indeed, some of the most exciting and practical uses of large databanks may be found precisely at the interface with those knowledge engineering tasks that have most confounded researchers in medical symbolic reasoning [5].

4 Mathematical Models of Physical Processes

4.1 Overview

Pathophysiologic processes can be well-described by mathematical formulae in a limited number of clinical problem areas. Such domains have lent themselves well to the development of computer-based decision aids since the issues are generally well-defined. The actual techniques used by such programs tend to reflect the details of the individual applications, the most celebrated of which have been in pharmacokinetics (specifically digitalis dosing), acid-base/electrolyte disorders, and respiratory care [63].

One or two cooperating experts in the field generally assist with the definition of pertinent variables and the mathematical characterization of the relationships among them. Often an interactive program is then developed which requests the relevant data, makes the appropriate computations, and provides a clinical analysis or recommendation for therapy based upon the computational results. Some of the programs have also involved branched-chain logic to guide decisions about what further data are needed for adequate analysis[5].

Programs to assist with digitalis dosing have progressed to the inclusion of broader medical knowledge over the last ten years. The earliest work was Jelliffe's [43] and was based upon his considerable experience studying the pharmacokinetics of the cardiac glycosides. His computer program used mathematical formulations based on parameters such as therapeutic goals (e.g., desired predicted blood levels), body weight, renal function, and route of administration. In one study he showed that computer recommendations reduced the frequency of adverse digitalis reactions from 35% to 12% [44]. Later, another group revised the Jelliffe model to permit a feedback loop in which the digitalis blood levels obtained with initial doses of the drug were considered

---

[5] "Branched-chain" logic refers to mechanisms by which portions of a decision network can be considered or ignored depending upon the data on a given case. For example, in an acid-base program the anion gap might be calculated and a branch-point could then determine whether the pathway for analyzing an elevated anion gap would be required. If the gap were not elevated, that whole portion of the logic network could be skipped.

in subsequent therapy recommendations [72], [89]. More recently, a third group in Boston, noting the insensitivity of the first two approaches to the kinds of nonnumeric observations that experts tend to use in modifying digitalis therapy, augmented the pharmacokinetic model with a patient-specific model of clinical status [31]. Running their system in a monitoring mode, in parallel with actual clinical practice on a cardiology service, they found that each patient in the trial in whom toxicity developed had received more digitalis than would have been recommended by their program.

4.2 Example

Perhaps the best known program in this category is the interactive system developed at Boston's Beth Israel Hospital by Bleich. Originally designed as a program for assessment of acid-base disorders [2], it was later expanded to consider electrolyte abnormalities as well [3], [4]. The knowledge in Bleich's program is a distillation of his own expertise regarding acid-base and electrolyte disorders. The system begins by collecting initial laboratory data from the physician seeking advice on a patient's management. Branched-chain logic is triggered by abnormalities in the initial data so that only the pertinent sections of the extensive decision pathways created by Bleich are explored. Essentially all questions asked by the program are numerical laboratory values or "yes-no" questions (e.g., "Does the patient have pitting edema?"). Depending upon the complexity and severity of the case, the program eventually generates an evaluation note that may vary in length from a few lines to several pages. Included are suggestions regarding possible causes of the observed abnormalities and suggestions for correcting them. Literature references are also provided.

Although the program was made available at several East Coast institutions, few physicians accepted it as an ongoing clinical tool. Bleich points out that part of the reason for this was the system's inherent educational impact; physicians simply began to anticipate its analysis after they had used it a few times [3]. More recently he has been experimenting with the program operating as a monitoring system[6], thereby avoiding direct interaction with the physician.

The system's lack of sustained acceptance by physicians is probably due to more than its educational impact, however. For example, there is no feedback in the system; every patient is seen as a new case and the program has no concept

[6]Personal communication with Dr. Bleich, 1975.

of following a patient's response to prior therapeutic measures. Furthermore, the program generates differential diagnosis lists but does not pursue specific etiologies; this can be particularly bothersome when there are multiple coexistent disturbances in a patient and the program simply suggests parallel lists of etiologies without noting or pursuing the possible interrelationships.

Finally, the system is highly individualized in that it contains consideration of specific relationships only when Bleich specifically thought to include them in the logic network. Of course human consultants also give personalized advice which may differ from that obtained from other experts. However, a group of researchers in Britain [79] who analyzed Bleich's program along with four other acid-base/electrolyte systems, found total agreement among the programs in only 20% of test cases when these systems were asked to define the acid-base disturbance and the degree of compensation present. Their analysis does not reveal which of the programs reached the correct decision, however, and it may be that the results are more an indictment of the other four programs than a valid criticism of the advice from Bleich's acid-base component.

4.3 Discussion of the Methodologies

The programs mentioned in this section are very different in several respects, and each tends to overlap with other methodologies we have discussed. Bleich's program, for example, is essentially a complicated clinical algorithm interfaced with mathematical formulations of electrolyte and acid-base pathophysiology. As such it suffers from the weaknesses of all algorithmic approaches, most importantly its highly structured and inflexible logic which is unable to contend with unforseen circumstances not specifically included in the algorithm. The digitalis dosing programs all draw on mathematical techniques from the field of biomedical modeling (not discussed here), but have recently shown more reliance on methods from other areas as well. In particular these have included symbolic reasoning methods that allow clinical expertise to be captured and utilized in conjunction with mathematical techniques [31]. The Boston group that developed this most recent digitalis program is interested in similarly developing an acid-base/electrolyte system so that judgmental knowledge of experts can be interfaced with the mathematical models of pathophysiology[7].

---

[7]Personal communication, 1978, with Prof. Peter Szolovits.

5 Statistical Pattern Matching Techniques

5.1 Overview

Pattern matching techniques define the mathematical relationship between measurable features and classifications of objects [12], [46]. In medicine, the presence or absence of each of several signs and symptoms in a patient may be definitive for the classification of the patient as "abnormal" or into the category of a specific disease. They are also used for prognosis [1], or predicting disease duration, time course, and outcomes. These techniques have been applied to a variety of medical domains, such as image processing and signal analysis, in addition to computer-assisted diagnosis.

In order to find the diagnostic pattern, or discriminant function, the method requires a training set of objects, for which the correct classification is already known, as well as reliable values for their measured features. If the form and parameters are not known for the statistical distributions underlying the features, then they must be estimated. Parametric techniques focus on learning the parameters of the probability density functions, while non-parametric (or "distribution-free") techniques make no assumptions about the form of the distributions. After training, then, the pattern can be matched to new, unclassified objects to aid in deciding the category to which the new object belongs[8].

There are numerous variations on this general methodology, most notably in the mathematical techniques used to extract characteristic measurements (the features) and to find and refine the pattern classifier during training. For example, linear regression analysis is a commonly used technique for finding the coefficients of an equation that defines a recurring pattern or category of diagnostic or prognostic interest. Recent work emphasizes structural relationships among sets of features more than statistical ones.

Three of the best known training criteria for the discriminant function are:

(a) Bayes' criterion: choose the function that has the minimum cost associated with incorrect diagnoses[9];

(b) clustering criterion: choose the function that produces the tightest clusters;

(c) least-squared-error criterion: choose the function that minimizes the squared differences between predicted and observed measurement values.

---

[8] It is possible to detect patterns, even without a known classification for objects in the training set, with so-called "unsupervised" learning techniques. Also, it is possible to work with both numerical and non-numerical measurements.

[9] See Section 6 for further discussion.

Ten commonly used mathematical models based on these criteria have been shown to produce remarkably similar diagnostic results for the same data [7].

5.2 Example

There are numerous papers reporting on the use of pattern recognition methods in medicine. Armitage [1] discusses three examples of prognostic studies, with an emphasis on regression methods. Siegel et al. [27] discuss uses of cluster analysis. One recent diagnostic application using Bayes' criterion [67] classifies patients having chest pains into three categories: D_1: acute myocardial infarction (MI); D_2: coronary insufficiency; and D_3: non-cardiac causes of chest pain. The need for early diagnosis of heart attacks without laboratory tests is a prevalent problem, yet physicians are known to misclassify about one third of the patients in categories D_1 and D_2 and about 80% of those in D_3. In order to determine the correct classification, each patient in the training set was classified after 3 days, based on laboratory data including electrocardiogram (ECG) and blood data (cardiac enzymes). There remained some uncertainty about several patients with "probable MI." Seventeen variables were selected from many: 9 features with continuous values (including age, heart rates, white blood count, and hemoglobin) and 8 features with discrete values (sex and 7 ECG features).

The training data were measurements on 247 patients. The decision rule was chosen using Bayes' theorem to compute the posterior probabilities of each diagnostic class given the feature vector X. ($X = [x_1, x_2, \ldots, x_{17}]$.[10]. Then a decision rule was chosen to minimize the probability of error, that is, to adjust the coefficients on the feature vector X [11] such that for the correct class D_i:

$$P(D_i|X)=MAX\ (P(D_1|X),\ P(D_2|X),\ P\ (D_3|X))$$

The class conditional probability density functions must be estimated initially, and the performance of the decision rule depends on the accuracy of the assumed model.

Using the same 247 patients for testing the approach, the trained

---

[10] The posterior probability of a diagnostic class, represented as $P(D_i|X)$, is the probability that a patient falls in diagnostic category D_i given that the feature vector X has been observed.

[11] See [56] for a study in which the coefficients are reported because of their medical import.

classifier averaged 80% correct diagnoses over the three classes, using only data available at the time of admission. Physicians, using more data than the computer, averaged only 50.5% correct over these three categories for the same patients. Training the classifier with a subset of the patients, and using the remainder for testing, produced nearly as good results.

5.3 Discussion of the Methodology

The number of reported medical applications of pattern recognition techniques is large, but there are also numerous problems associated with the methodology. The most obvious difficulties are choosing the set of features in the first place, collecting reliable measurements on a large sample, and verifying the initial classifications among the training data. Current techniques are inadequate for problems in which trends or movement of features are important characteristics of the categories. Also the problems for which existing techniques are accurate are those that are well characterized by a small number of features ("dimensions of the space").

As with all techniques based on statistics, the size of the sample used to define the categories is an important consideration. As the number of important features and the number of relevant categories increase, the required size of the training set also increases. In one test [7], pattern classifiers trained to discriminate among 20 disease categories from 50 symptoms were correct 51% - 64% of the time. The same methods were used to train classifiers to discriminate between 2 of the diseases, from the same 50 symptoms, and produced correct diagnoses 92% - 98% of the time.

The context in which a local pattern is identified raises problems related to the issue of utilizing medical knowledge. It is difficult to find and use classifiers that are best for a small decision, such as whether an area of an X-ray is inside or outside the heart, and integrate those into a global classifier, such as one for abnormal heart volume.

Accurate application of a classifier in a hospital setting also requires that the measurements in that clinical environment are consistent with the measurements used to train the classifier initially. For example, if diseases and symptoms are defined differently in the new setting, or if lab test values are reported in different ranges — or different lab tests used -- then decisions based on the classification are not reliable.

Pattern recognition techniques are often misapplied in medical domains in

which the assumptions are violated. Some of the difficulties noted above are avoided in systems that integrate structural knowledge into the numerical methods and in systems that integrate human and machine capabilities into single, interactive systems. These modifications will overcome one of the major difficulties seen in completely automated systems, that of providing the system with good "intuitions" based on an expert's a priori knowledge and experience [46].

6 Bayesian Statistical Approaches

6.1 Overview

More work has been done on Bayesian approaches to computer-based medical decision making than on any of the other methodologies we have discussed. The appeal of Bayes' Theorem [12] is clear: it potentially offers an exact method for computing the probability of a disease based on observations and data regarding the frequency with which these observations are known to occur for specified diseases. In several domains the technique has been shown to be exceedingly accurate, but there are also several limitations to the approach which we discuss below.

In its simplest formulation, Bayes' Theorem can be seen as a mechanism to calculate the probability of a disease, in light of specified evidence, from the a priori probability of the disease and the conditional probabilities relating the observations to the diseases in which they may occur. For example, suppose disease D_i is one of n mutually exclusive diagnoses under consideration and E is the evidence or observations supporting that diagnosis. Then if $P(D_i)$ is the a priori probability of the ith disease:

$$P(D_i|E) = \frac{P(D_i)\ P(E|D_i)}{\sum_{j=1}^{n} P(D_j)\ P(E|D_j)}$$

The theorem can also be represented or derived in a variety of other forms, including an odds/likelihood ratio formulation. We cannot include such details here, but any introductory statistics book or Lusted's classic volume [58] presents the subject in considerable detail.

---

[12] also often referred to as Bayes' rule, discriminant, or criterion

Among the most commonly recognized problems with the utilization of a Bayesian approach is the large amount of data required to determine all the conditional probabilities needed in the rigorous application of the formula. Chart review or computer-based analysis of large databanks occasionally allows most of the necessary conditional probabilities to be obtained. A variety of additional assumptions must be made. For example: (1) the diseases under consideration are assumed mutually exclusive and exhaustive (i.e., the patient is assumed to have one of the n diseases, (2) the clinical observations are assumed to be conditionally independent over a given disease[13], and (3) the incidence of the symptoms of a disease is assumed to be stationary (i.e., the model generally does not allow for changes in disease patterns over time).

One of the earliest Bayesian programs was Warner's system for the diagnosis of congenital heart disease [107]. He compiled data on 83 patients and generated a symptom-disease matrix consisting of 53 symptoms (attributes) and 35 disease entities. The diagnostic performance of the computer, based on the presence or absence of the 53 symptoms in a new patient, was then compared to that of two experienced physicians. The program was shown to reach diagnoses with an accuracy equal to that of the experts. Furthermore, system performance was shown to improve as the statistics in the symptom-disease matrix stabilized with the addition of increasing numbers of patients.

In 1968 Gorry and Barnett pointed out that Warner's program had required making all 53 observations for every patient to be diagnosed, a situation which would not be realistic for many clinical applications. They therefore utilized a modification of Bayes' Theorem in which observations are considered sequentially. Their computer program analyzed observations one at a time, suggested which test would be most useful if performed next, and included termination criteria so that a diagnosis could be reached, when appropriate, without needing to make all the observations [28]. Decisions regarding tests and termination were made on the basis of calculations of expected costs and benefits at each step in the logical process[14]. Using the same symptom-disease matrix developed by Warner, they were able to attain equivalent diagnostic

---

[13]The purest form of Bayes' Theorem allows conditional dependencies, and the order in which evidence is obtained, to be explicitly considered in the analysis. However, the number of required conditional probabilities is so unwieldy that conditional independence of observations, and non-dependence on the order of observations, is generally assumed [101].

[14]See the decision theory discussion in Section 7.

performance using only 6.9 tests on average[15]. They pointed out that, because the costs of medical tests may be significant (in terms of patient discomfort, time expended, and financial expense), the use of inefficient testing sequences should be regarded as ineffective diagnosis. Warner has also more recently included Gorry and Barnett's sequential diagnosis approach in an application regarding structured patient history-taking [110].

The medical computing literature now includes many examples of Bayesian diagnosis programs, most of which have used the nonsequential approach, in addition to the necessary assumptions of symptom independence and mutual exclusivity of disease as discussed above. One particularly successful research effort has been chosen for discussion.

6.2 Example

Since the late 1960's deDombal and associates, at the University of Leeds in England, have been studying the diagnostic process and developing computer-based decision aids using Bayesian probability theory. Their area of investigation has been gastrointestinal diseases, originally acute abdominal pain [10] with more recent analyses of dyspepsia [39] and gastric carcinoma [125].

Their program for assessment of acute abdominal pain was evaluated in the emergency room of their affiliated hospital [10]. Emergency physicians filled out data sheets summarizing clinical and laboratory findings on 304 patients presenting with abdominal pain of acute onset. The data from these sheets became the attributes that were subjected to Bayesian analysis; the required conditional probabilities had been previously compiled from a large group of patients with one of 7 possible diagnoses[16]. Thus the Bayesian formulation assumed each patient had one of these diseases and would select the most likely on the basis of recorded observations. Diagnostic suggestions were obtained in batch mode and did not require direct interaction between physician and computer; the program could generate results in from 30 seconds to 15 minutes depending upon the level of system use at the time of analysis [38]. Thus the computer output could have been made available to the emergency room physician, on average, within 5 minutes after the data form was completed and handed to the technician assisting with the study.

---

[15] Tests for determining attributes were defined somewhat differently than they had been by Warner. Thus the maximum number of tests was 31 rather than the 53 observations used in the original study.

[16] appendicitis, diverticulitis, perforated ulcer, cholecystitis, small bowel obstruction, pancreatitis, and non-specific abdominal pain.

During the study [10], however, these computer-generated diagnoses were simply saved and later compared to (a) the diagnoses reached by the attending clinicians, and (b) the ultimate diagnosis verified at surgery or through appropriate tests. Although the clinicians reached the correct diagnosis in only 65%-80% of the 304 cases (with accuracy depending upon the individual's training and experience), the program was correct in 91.8% of cases. Furthermore, in 6 of the 7 disease categories the computer was proved more likely than the senior clinician in charge of a case to assign the patient to the correct disease category. Of particular interest was the program's accuracy regarding appendicitis - a diagnosis which is often made incorrectly. In no cases of appendicitis did the computer fail to make the correct diagnosis, and in only six cases were patients with non-specific abdominal pain incorrectly classified as having appendicitis. Based on the actual clinical decisions, however, over 20 patients with non-specific abdominal pain were unnecessarily taken to surgery for appendicitis, and in six cases patients with appendicitis were "watched" for over eight hours before they were finally taken to the operating room.

These investigators also performed a fascinating experiment in which they compared the program's performance based on data derived from 600 real patients, with the accuracy the system achieved using "estimates" of conditional probabilities obtained from experts [54][17]. As discussed above, the program was significantly more effective than the unaided clinician when real-life data were utilized. However, it performed significantly less well than clinicians when expert estimates were used. The results supported what several other observers have found, namely that physicians often have very little idea of the "true" probabilities for symptom-disease relationships.

Another Leeds study of note was an analysis of the effect of the system on the performance of clinicians [11]. The trial we have mentioned that involved 304 patients was eventually extended to 552 before termination. Although the computer's accuracy remained in the range of 91% throughout this period, the performance of clinicians was noted to improve markedly over time. Fewer negative laparotomies were performed, for example, and the number of acute appendices that perforated (ruptured) also declined. However, these data reverted to baseline after the study was terminated, suggesting that the

---

[17] Such estimates are referred to as "subjective" or "personal" probabilities, and some investigators have argued that they should be utilized in Bayesian systems when formally derived conditional probabilities are not available [58].

constant awareness of computer monitoring and feedback regarding system performance had temporarily generated a heightened awareness of intellectual processes among the hospital's surgeons.

6.3 Discussion of the Methodology

The ideal matching of the problem of acute abdominal pain and Bayesian analysis must also be emphasized; the methodology cannot necessarily be as effectively applied in other medical domains where the following limitations of the Bayesian approach may have a greater impact.

(1) The assumption of conditional independence of symptoms usually does not apply and can lead to substantial errors in certain settings [66]. This has led some investigators to seek new numerical techniques that avoid the independence assumption [8]. If a pure Bayesian formulation is utilized without making the independence assumption, however, the number of required conditional probabilities becomes prohibitive for complex real world problems [101].

(2) The assumption of mutual exclusivity and exhaustiveness of disease categories is usually false. In actual practice concurrent and overlapping disease categories are common. In deDombal's system, for example, many of the abdominal pain diagnoses missed were outside the seven "recognized" possibilities; if a program starts with an assumption that it need only consider a small number of defined likely diagnoses, it will inevitably miss the rare or unexpected cases - precisely the ones with which the clinician is most apt to need assistance.

(3) In many domains it may be inaccurate to assume that relevant conditional probabilities are stable over time (e.g., the likelihood that a particular bacterium will be sensitive to a specific antibiotic). Furthermore, diagnostic categories and definitions are constantly changing, as are physicians' observational techniques, thereby invalidating data previously accumulated. A similar problem results from variations in *a priori* probabilities depending upon the population from which a patient is drawn. Some observers feel that these are major limitations to the use of Bayesian techniques [13].

In general, then, a purely Bayesian approach can so constrain problem formulation as to make a particular application unrealistic and hence unworkable. Furthermore, even when diagnostic performance is excellent such as in deDombal's approach to abdominal pain evaluation, clinical implementation and system acceptance will generally be difficult.

7 Decision Theoretical Approaches

7.1 Overview

Bayes' Theorem is only one of several techniques used in the larger field of decision analysis, and there has recently been increasing interest in the ways in which decision theory might be applied to medicine and adapted for automation. Several excellent reviews of the field are available in basic reviews [40], textbooks [78], and medically-oriented journal articles [61], [87], [102]. In general terms, decision analysis can be seen as any attempt to consider values associated with choices, as well as probabilities, in order to analyze the processes by which decisions are made or should be made. Schwartz identifies the calculation of "expected value" as central to formal decision analysis [87]. Ginsberg contrasts medical classification problems (e.g., diagnosis) with broader decision problems (e.g., "What should I do for this patient?"), and asserts that most important medical decisions fall in the latter category and are best approached through decision analysis [25]. The following topics are among the central issues in the field.

(1) Decision Trees. The decision making process can be seen as a sequence of steps in which the clinician selects a path through a network of plausible events and actions. Nodes in this tree-shaped network are of two kinds: decision nodes, where the clinician must choose from a set of actions, and chance nodes, where the outcome is not directly controlled by the clinician but is a probabilistic response of the patient to some action taken. For example, a physician may choose to perform a certain test (decision node) but the occurrence or nonoccurrence of complications may be largely a matter of statistical likelihood (chance node). By analyzing a difficult decision process before taking any actions, it may be possible to delineate in advance all pertinent chance and decision nodes, all plausible outcomes, plus the paths by which these outcomes might be reached. Furthermore, data may exist to allow specific probabilities to be associated with each chance node in the tree.

(2) Expected Values. In actual practice physicians make sequential decisions based on more than the probabilities associated with the chance node that follows. For example, the best possible outcome is not necessarily sought if the costs associated with that "path" far outweigh those along alternate pathways (e.g., a definitive diagnosis may not be sought if the required testing procedure is expensive or painful and patient management will be unaffected;

similarly, some patients prefer to "live with" an inquinal hernia rather than undergo a surgical repair procedure). Thus anticipated "costs" (financial, complications, discomfort, patient preference) can be associated with the decision nodes. Using the probabilities at chance nodes, the costs at decision nodes, and the "value" of the various outcomes, an "expected value" for each pathway through the tree (and in turn each node) can be calculated. The ideal pathway, then, is the one which maximizes the expected value.

(3) Eliciting Values. Obtaining from physicians and patients the cost and values they associate with various tests and outcomes can be a formidable problem, particularly since formal analysis requires expressing the various costs in standardized units. One approach has been simply to ask for value ratings on a hypothetical scale, but it can be difficult to get the physician or patient to keep the values[18] separate from their knowledge of the probabilities linked to the associated chance nodes. An alternate approach has been the development of lottery games. Inferences regarding values can be made by identifying the odds, in a hypothetical lottery, at which the physician or patient is indifferent regarding taking a course of action with certain outcome and betting on a course with preferable outcome but with a finite chance of significant negative costs if the "bet" is lost. In certain settings this approach may be accepted and provide important guidelines in decision making [71].

(4) Test Evaluation. Since the tests which lie at decision nodes are central to clinical decision analysis, it is crucial to know the predictive value of tests that are available. This leads to consideration of test sensitivity, specificity, receiver operator characteristic curves, and sensitivity analysis. Such issues are discussed by Komaroff et al. in this issue of the PROCEEDINGS and have also been summarized elsewhere in the clinical literature [62].

Many of the major studies of clinical decision analysis have not specifically involved computer implementations. Schwartz et al. examined the workup of renal vascular hypertension, developing arguments to show that for certain kinds of cases a purely qualitative theoretical approach was feasible and useful [87]. However, they showed that for more complex clinically challenging cases the decisions could not be adequately sorted out without the introduction of numerical techniques. Since it was impractical to assume that

---

[18] also termed "utilities" in some references; hence the term "utility theory" [78].

clinicians would ever take the time to carry out a detailed quantitative decision analysis by hand, they pointed out the logical role for the computer in assisting with such tasks and accordingly developed the system we discuss as an example below [29].

Other colleagues of Schwartz at Tufts have been similarly active in applying decision theory to clinical problems. Pauker and Kassirer have examined applications of formal cost-benefit analysis to therapy selection [68] and Pauker has also looked at possible applications of the theory to the management of patients with coronary artery disease [70]. An entire issue of the New England Journal of Medicine has also been devoted to papers on this methodology [41].

7.2 Example

Computer implementations of clinical decision analysis have appeared with increasing frequency since the mid-1960's. Perhaps the earliest major work was that of Ginsberg at Rand Corporation [24], with more recent systems reported by Pliskin and Beck [74] and Safran et al. [85].

We will briefly describe here the program of Gorry et al., developed for the management of acute renal failure [29]. Drawing upon Gorry's experience with the sequential Bayesian approach previously mentioned [28], the investigators recognized the need to incorporate some way of balancing the dangers and discomforts of a procedure against the value of the information to be gained. They divided their program into two parts: phase I considered only tests with minimal risk (e.g., history, examination, blood tests) and phase II considered procedures involving more risk and inconvenience. The phase I program considered 14 of the most common causes of renal failure and utilized a sequential test selection process based on Bayes' Theorem and omitting more advanced decision theoretical methodology [28]. The conditional probabilities utilized were subjective estimates obtained from an expert nephrologist and were therefore potentially as problematic as those discussed by Leaper et al. [54] (see Section 6.2). The researchers found that they had no choice but to use expert estimates, however, since detailed quantitative data were not available either in databanks nor the literature.

It is in the phase II program that the methods of decision theory were employed because it was in this portion of the decision process that the risks of procedures became important considerations. At each step in the decision

process this program considers whether it is best to treat the patient immediately or to first carry out an additional diagnostic test. To make this decision the program identifies the treatment with the highest current expected value (in the absence of further testing), and compares this with the expected values of treatments that could be instituted if another diagnostic test were performed. Comparison of the expected values are made in light of the risk of the test in order to determine whether the overall expected value of the test is greater than that of immediate treatment. The relevant values and probabilities of outcomes of treatment were obtained as subjective estimates from nephrologists in the same way that symptom-disease data had been obtained. All estimates were gradually refined as they gained experience using the program, however.

The program was evaluated on 18 test cases in which the true diagnosis was uncertain but two expert nephrologists were willing to make management decisions. In 14 of the cases the program selected the same therapeutic plan or diagnostic test as was chosen by the experts. For three of the four remaining cases the program's decision was the physicians' second choice and was, they felt, a reasonable alternative plan of action. In the last case the physicians also accepted the program's decision as reasonable although it was not among their first two choices.

7.3 Discussion of the Methodology

The excellent performance of Gorry's program, despite its reliance on subjective estimates from experts, may serve to emphasize the importance of the clinical analysis that underlies the decision theoretical approach. The reasoning steps in managing clinical cases have been dissected in such detail that small errors in the probability estimates are apparently much less important than they were for deDombal's purely Bayesian approach [54]. Gorry suggests this may be simply because the decisions made by the program are based on the combination of large aggregates of such numbers, but this argument should apply equally for a Bayesian system. It seems to us more likely that distillation of the clinical domain in a formal decision tree gives the program so much more knowledge of the clinical problem that the quantitative details become somewhat less critical to overall system operation. The explicit decision network is a powerful knowledge structure; the "knowledge" in deDombal's system lies in conditional probabilities alone and there is no larger

scheme to override the propagation of error as these probabilities are mathematically manipulated by the Bayesian routines.

The decision theory approach is not without problems, however. Perhaps the most difficult problem is assigning numerical values (e.g., dollars) to a human life or a day of health, etc. Some critics feel this is a major limitation to the methodology [112]. Overlapping or coincident diseases are also not well-managed, unless specifically included in the analysis, and the Bayesian foundation for many of the calculations still assumes mutually exclusive and exhaustive disease categories. Problems of symptom conditional dependence still remain, and there is no easy way to include knowledge regarding the time course of diseases. Gorry points out that his program was also incapable of recognizing circumstances in which two or more actions should be carried out concurrently. Furthermore decision theory per se does not provide the kind of focusing mechanisms that clinicians tend to use when they assume an initial diagnostic hypothesis in dealing with a patient and discard it only if subsequent data make that hypothesis no longer tenable. Other similar strategies of clinical reasoning are becoming increasingly well-recognized [48] and account in large part for the applications of symbolic reasoning techniques to be discussed in the next section.

8 Symbolic Reasoning Approaches

8.1 Overview

In the early 1970's researchers at several institutions simultaneously began to investigate the potential applications to clinical decision making of symbolic reasoning techniques drawn from the branch of computer science known as artificial intelligence (AI). The field is well-reviewed in a recent book by Winston [120]. Although the term "artificial intelligence" has never been uniformly defined, it is generally accepted to include those computer applications in which the tasks require largely symbolic inference rather than numeric calculation. Examples include programs that reason about mineral exploration, organic chemistry, or molecular biology; programs that converse in English and understand spoken sentences; and programs that generate theories from observations.

Such programs gain their power from qualitative, experimental judgments - codified in so-called "rules-of-thumb" or "heuristics" - in contrast to

numerical calculation programs whose power derives from the analytical equations used. The heuristics focus the attention of the reasoning program on parts of the problem that seem most critical and parts of the knowledge base that seem most relevant. They also guide the application of the domain knowledge to an individual case by deleting items from consideration as well as focusing on items. The result is that these programs pursue a line of reasoning as opposed to following a sequence of steps in a calculation. Among the earliest symbolic inference programs in medicine was the diagnostic interviewing system of Kleinmuntz [49]. Other early work included Wortman's information processing system, the performance of which was largely motivated by a desire to understand and simulate the psychological processes of neurologists reaching diagnoses [121].

It was a landmark paper by Gorry in 1973, however, that first critically analyzed conventional approaches to computer-based clinical decision making and outlined his motivation for turning to newer symbolic techniques [30]. He used the acute renal failure program discussed in Section 7.2 [29] as an example of the problems arising when decision analysis is used alone. In particular, he analyzed some of the cases on which the renal failure program had failed but the physicians considering the cases had performed well. His conclusions from these observations include the following four points.

(1) Clinical judgment is based less on detailed knowledge of pathophysiology than it is on gross chunks of knowledge and a good deal of detailed experience from which rules of thumb are derived.

(2) Clinicians know facts, of course, but their knowledge is also largely judgmental. The rules they learn allow them to focus attention and generate hypotheses quickly. Such heuristics permit them to avoid detailed search through the entire problem space.

(3) Clinicians recognize levels of belief or certainty associated with many of the rules they use, but they do not routinely quantitate or utilize these certainty concepts in any formal statistical manner.

(4) It is easier for experts to state their rules in response to perceived misconceptions in others than it is for them to generate such decision criteria a priori.

In the renal failure program medical knowledge had been embedded in the structure of the decision tree. This knowledge was never explicit, and additions to the experts' judgmental rules had generally required changes to the tree itself.

Based on observations such as those above, Gorry identified at least three important problems for investigation:

(1) Concept Formation. Clinical decision aids had traditionally had no true "understanding" of medicine. Although explicit decision trees had given the decision theory programs a greater sense of the pertinent associations, medical knowledge and the heuristics for problem solving in the field had never been explicitly represented nor utilized. So-called "common sense" was often clearly lacking when the programs failed, and this was often what most alienated potential physician users.

(2) Language Development. Both for capturing knowledge from collaborating experts, and for communicating with physician users, Gorry argued that further research on the development of computer-based linguistic capabilities was crucial.

(3) Explanation. Diagnostic programs had seldom emphasized an ability to explain the basis for their decisions in terms understandable to the physician. System acceptability was therefore inevitably limited; the physician would often have no basis for deciding whether to accept the program's advice, and might therefore resent what could be perceived as an attempt to dictate the practice of medicine.

Gorry's group at MIT and Tufts developed new approaches to examining the renal failure problem in light of these observations [69].

Due to the limitations of the older techniques, it was perhaps inevitable that some medical researchers would turn to the AI field for new methodologies. Major research areas in AI include knowledge representation, heuristic search, natural language understanding and generation, and models of thought processes — all topics clearly pertinent to the problems we have been discussing. Furthermore, AI researchers were beginning to look for applications to which they could apply some of the techniques they had developed in theoretical domains. This community of researchers has grown in recent years, and a recent issue of Artificial Intelligence was devoted entirely to applications of AI to biology, medicine and chemistry [98][19].

---

[19] Many of the systems described in this issue were developed on the SUMEX-AIM computing resource, a nationally shared system devoted entirely to applications of AI to the biomedical sciences. The SUMEX-AIM computer is physically located at Stanford University but is used by researchers nationwide via connections to the TYMNET. The resource is funded by the Division of Research Resources, Biotechnology Branch, National Institutes of Health.

Among the programs using symbolic reasoning techniques are several systems that have been particularly novel and successful. Pople and Myers have developed a system called INTERNIST that assists with test selection for the diagnosis of all diseases in internal medicine [75]. This awesome task has been remarkably successful to date, with the program correctly diagnosing a large percentage of the complex cases selected from clinical pathologic conferences in the major medical journals[2C]. The program utilizes a hierarchic disease categorization, an ad hoc scoring system for quantifying symptom-disease relationships, plus some clever heuristics for focusing attention, discriminating between competing hypotheses, and diagnosing concurrent diseases [76]. The system currently has an inadequate human interface, however, and is not yet implemented for clinical trials.

At Rutgers University Weiss', Kulikowski, and Safir have developed a model of ophthomologic reasoning regarding disease processes in the eye, specifically glaucoma [117]. In this specialized application area it has been possible to map relationships between observations, pathophysiologic states, and disease categories. The resulting causal associational network (termed CASNET) forms the basis for a reasoning program that gives advice regarding disease states in glaucoma patients and generates management recommendations.

For the AI researchers the question of how best to manage uncertainty in medical reasoning remains a central issue. All the programs mentioned have developed ad hoc weighting programs and avoided formal statistical approaches. Others have turned to the work of statisticians and philosophers of science who have devised theories of approximate or inexact reasoning. For example, Wechsler [114] describes a program that is based upon Zadeh's fuzzy set theory [124]. Shortliffe and Buchanan [94] have turned to confirmation theory for their model of inexact reasoning in medicine.

8.2 Example

The symbolic reasoning program selected for discussion is the MYCIN System at Stanford University [95]. The researchers cited a variety of design considerations which motivated the selection of AI methodologies for the consultation system they were developing [92]. They primarily wanted it to be useful to physicians and therefore emphasized the selection of a problem domain in which physicians had been shown to err frequently, namely the selection of

---

[2C]Data communicated by Drs. Pople and Myers at the Second Annual A.I.M. Workshop, Rutgers University, June 1976.

antibiotics for patients with infections. They also cited human issues that they felt were crucial to make the system acceptable to physicians:

(1) it should be able to explain its decisions in terms a line of reasoning that a physician can understand;

(2) it should be able to justify its performance by responding to questions expressed in simple English;

(3) it should be able to "learn" new information rapidly by interacting directly with experts;

(4) its knowledge should be easily modifiable so that perceived errors can be corrected rapidly before they recur in another case; and

(5) the interaction should be engineered with the user in mind (in terms of prompts, answers, and information volunteered by the system as well as by the users).

All these design goals were based on the observation that previous computer decision aids had generally been poorly accepted by physicians, even when they were shown to perform well on the tasks for which they were designed. MYCIN's developers felt that barriers to acceptance were largely conceptual and could be counteracted in large part if a system were perceived as a clinical tool rather than a dogmatic replacement for the primary physician's own reasoning.

Knowledge of infectious diseases is represented in MYCIN as production rules, each containing a "packet" of knowledge obtained from collaborating experts [95][21]. A production rule is simply a conditional statement which relates observations to associated inferences that may be drawn. For example, a MYCIN rule might state that "if a bacterium is a gram positive coccus growing in chains, then it is apt to be a streptococcus." MYCIN's power is derived from such rules in a variety of ways:

(1) it is the program that determines which rules to use and how they should be chained together to make decisions about a specific case[22];

(2) the rules can be stored in a machine-readable format but translated into English for display to physicians;

(3) by removing, altering, or adding rules, the system's knowledge structures can be rapidly modified without explicitly restructuring the entire knowledge base; and

(4) the rules themselves can often form a coherent explanation of system reasoning if the relevant ones are translated into English and displayed in response to a user's question.

Associated with all rules and inferences are numerical weights reflecting the degree of certainty associated with them. These numbers, termed certainty factors, form the basis for the system's inexact reasoning in this complex task

---

[21]Production rules are a methodology frequently employed in AI research [9] and effectively applied to other scientific problem domains [6].

[22]The control structure utilized is termed "goal-oriented" and is similar to the consequent-theorem methodology used in Hewitt's PLANNER [37].

domain [94]. They allow the judgmental knowledge of experts to be captured in rule form and then utilized in a consistent fashion.

The MYCIN System has been evaluated regarding its performance at therapy selection for patients with either septicemia [123] or meningitis [122]. The program performs comparably with experts in these two task domains, but as yet it has no rules regarding the other infectious disease problem areas. Further knowledge base development will therefore be required before MYCIN is made available for clinical use; hence questions regarding its acceptability to physicians cannot yet be assessed. However, the required implementation stages have been delineated [93], attention has been paid to all the design criteria mentioned above, and the program does have a powerful explanation capability [88].

8.3 Discussion of the Methodology

Symbolic reasoning techniques differ from the other methodologies mentioned in this article in that the computer techniques themselves are as yet experimental and rapidly changing. Whereas the computations involved in Bayes' Theorem, for example, involve straightforward application of computing techniques already well-developed, basic researchers in computer science continue to develop new methodologies for knowledge representation, language understanding, heuristic search, and the other symbolic reasoning problems we have mentioned. Thus the AI programs tend to be developed in highly experimental environments where short term practical results are often unlikely to be found. The programs typically require large amounts of space and tend to be slow, particularly in time-sharing environments. As has been true for most of the methodologies discussed, AI researchers have still not developed adequate methods for handling concurrent diseases, assessing the time course of disease, nor acquiring adequate structured knowledge from experts. Furthermore, inexact reasoning techniques tend to be developed and justified largely on intuitive grounds.

Despite these significant limitations, the techniques of artificial intelligence do provide a way to respond to many of Gorry's observations regarding the inadequacies of prior methodologies as described above [30]. There are now several programs responsive to his criticisms. Szolovits and Pauker have recently reviewed some applications of AI to medicine and have attempted to weigh the successes of this young field against the very real